Deep Learning (CSE641/ECE555)
Quiz-3 (15 Marks) (Duration 60 min)

Name .....
Roll No. .....

Question 1-6 [1 Marks], Question 7-9 [2 Marks], Question 10 [3 marks]

Beginner

1. Ambiguous queries with polysemous terms (e.g., apple: fruit vs. company) degrade search accuracy. Which methods (select all that apply) most effectively mitigate this issue in search engine?

   (a) Word sense disambiguation via contextual embeddings
   (b) Aggressive stemming of query terms
   (c) Hybrid retrieval: keyword + vector-based search
   (d) Expanding the indexed document corpus
   (e) Query expansion using static synonym dictionaries

   A, C. Implementing word sense disambiguation using contextual embeddings and deploying hybrid search combining BM25 and vector retrieval would most effectively improve retrieval accuracy in this scenario.

2. A law firm requires an AI assistant that provides accurate legal citations with verifiable sources. Which method(s) support reliable attribution?

   a.)Pretraining on 10,000 high-quality legal textbooks b.)Hybrid retrieval (vector DB + keyword matching)
   c.)Post-processing with rule-based citation validation d.)Using a smaller LLM to reduce hallucination risk

   B, C. Hybrid retrieval systems and rule-based citation validators help ensure that references to case laws are traceable and accurate.

3. How does Chain-of-Thought (CoT) prompting improve hallucination control in Retrieval-Augmented Generation (RAG) systems?

   (a) Increases parameter utilization via added computation
   (b) Introduces noise to promote conservative outputs
   (c) Integrates retrieved content through structured reasoning
   (d) Replaces retrieval with internal reasoning mechanisms

   C. Chain of Thought (CoT) prompting enhances RAG systems by enabling the model to integrate retrieved information into structured reasoning steps, reducing the likelihood of hallucinations.

4. In RAG-based systems aimed at minimizing hallucinations, which factor can counterintuitively increase hallucination risk?

   (a) Overreliance on pre-trained weights without task-specific fine-tuning
   (b) Retrieval errors due to semantic mismatch in vector search
   (c) Large context windows leading to attention dispersion
   (d) Lack of human-in-the-loop feedback during generation

   B. Data mix-up, where semantic search retrieves incorrect or irrelevant information, can paradoxically increase hallucinations in RAG systems.

5. You're designing a QA system over 50-page technical manuals and considering two approaches:

   - A standard-context LLM (8k tokens) with an advanced RAG pipeline (semantic chunking, re-ranking, summarization)
   - A long-context LLM (128k tokens) that can process nearly the full manual in a single pass.

   Which statement best captures the trade-offs related to context window limitations?

   (a) Long-context LLMs offer superior fidelity for tasks requiring global context, as they can directly attend across the entire input.

   (b) Standard LLMs with RAG typically involve less complex infrastructure and exhibit more stable inference costs than long-context models.

   (c) Long-context models may underutilize extended input due to attention degradation and are more resource-intensive per query.

   (d) RAG pipelines may falter on tasks needing cross-section synthesis, where long-context models inherently excel by processing the full document.

   C. The long-context LLM (Approach 2) may suffer from reduced attention to information in the middle of its vast context (https://arxiv.org/pdf/2307.03172) and incurs higher computational costs per query, potentially offsetting its benefits over a well-optimized RAG system (Approach 1) for certain tasks.

6. A neural network layer has activations with mean 2.5, range [-1.0, 10.0], and most values concentrated in [1.0, 4.0]. When quantizing to INT8, which method is more likely to preserve information effectively?

   (a) Symmetric quantization — preserves zero-point alignment and simplifies hardware operations

   (b) Asymmetric quantization — better maps skewed distributions by utilizing the full dynamic range

   (c) Symmetric quantization — more effectively captures the extended range [-10.0, 10.0]

   (d) Both perform similarly — INT8 range is sufficient to encode the data

   B. Asymmetric quantization would likely preserve more information for this distribution. The activation values are highly skewed (range [-1.0, 10.0] with most values between [1.0, 4.0]), meaning there's a significant imbalance between negative and positive values. Asymmetric quantization introduces a zero-point offset that allows better utilization of the available quantization range for such skewed distributions.

Intermediate

7. Consider a deep neural network with L layers where each layer performs the operation $h_l = \phi(W_l \cdot q_s(h_{l-1}) + b_l)$, where ϕ is a Lipschitz continuous activation function with constant K_ϕ, W_l is the weight matrix, b_l is the bias vector, and q_s is the stochastic float8 quantization function defined by:

$$q_s(x) = \begin{cases} q_i & \text{with probability } 1 - p(x) \\ q_{i+1} & \text{with probability } p(x) \end{cases}$$

   where q_i and q_{i+1} are consecutive representable float8 values surrounding x, and $p(x) = \frac{x - q_i}{q_{i+1} - q_i}$. Derive the exact expression for the variance of the quantization error at each layer: $\text{Var}[q_s(h_{l-1}) - h_{l-1}]$

   For a value x between two consecutive float8 representable values q_i and q_{i+1}, the quantization error using stochastic rounding follows a Bernoulli distribution:

$$e_q(x) = q_s(x) - x = \begin{cases} q_i - x & \text{with probability } 1 - p(x) \\ q_{i+1} - x & \text{with probability } p(x) \end{cases}$$

   The variance of this quantization error is:

$$\text{Var}[e_q(x)] = E[(e_q(x))^2] - (E[e_q(x)])^2$$

   Since our stochastic rounding scheme is unbiased, $E[e_q(x)] = 0$. Therefore:

$$\text{Var}[e_q(x)] = E[(e_q(x))^2] = p(x)(q_{i+1} - x)^2 + (1 - p(x))(q_i - x)^2$$

Substituting $p(x) = \frac{x-q_i}{q_{i+1}-q_i}$:

$$\begin{aligned}\mathrm{Var}[e_q(x)] &= \frac{x-q_i}{q_{i+1}-q_i}(q_{i+1}-x)^2 + \frac{q_{i+1}-x}{q_{i+1}-q_i}(q_i-x)^2 \\ &= \frac{(x-q_i)(q_{i+1}-x)^2 + (q_{i+1}-x)(q_i-x)^2}{q_{i+1}-q_i}\end{aligned}$$

After algebraic simplification:

$$\begin{aligned}\mathrm{Var}[e_q(x)] &= \frac{(x-q_i)(q_{i+1}-x)(q_{i+1}-q_i)}{q_{i+1}-q_i} \\ &= (x-q_i)(q_{i+1}-x)\end{aligned}$$

The maximum variance occurs at $x = \frac{q_i+q_{i+1}}{2}$ (midpoint between representable values), yielding:

$$\mathrm{Var}_{\max}[e_q(x)] = \frac{(q_{i+1}-q_i)^2}{4}$$

For float8 with 4 mantissa bits, consecutive normalized values differ by 2^{-4} relative to the exponent, so the maximum variance is 2^{-10} times the square of the exponent scale.

8. which component is most likely responsible for accuracy degradation when using the following mixed-precision quantization in a transformer model and why?

   (a) Embedding Layers: 8-bit
   (b) Self-Attention Weights: 4-bit
   (c) Feed-Forward Layers: 8-bit
   (d) Layer Normalization: 16-bit

   B. Quantizing self-attention weights to 4-bit introduces coarse granularity and high quantization error, which degrades accuracy due to the precision-sensitive nature of attention computations like dot-products and projection matrices.

9. Select all applicable techniques for each LLM limitation:

LLM Limitation	**A**	**B**	**C**	**D**	**E**	**F**
Hallucination						
Knowledge Cutoff						
Chat Personalization						
Limited Context Length						

Legend – Techniques (Column Letters):

A.) RAG B.)Google Search Tool Use C.)Chain of Thought D.)Source Attribution
E.)Reasoning Agents F.)Memory Management

1. Hallucination → A, C, D, E
2. Knowledge Cutoff → A, B
3. Chat Personalization → F
4. Limited Context Length → A, B, F

Advanced

10. In loss landscape visualization, filter normalization is defined as:

$$\tilde{d} = \frac{d}{\|d\|_{\text{filter}}} \cdot \|\theta\|_{\text{filter}}$$

where the filter-wise norm for a direction vector is:

$$\|d\|_{\text{filter}} = \sqrt{\sum_{l=1}^{L}\sum_{j=1}^{n_l} \frac{\|d_j^l\|_F^2}{\|\theta_j^l\|_F^2} \cdot \|\theta_j^l\|_F^2}$$

with θ_j^l representing filter j in layer l, and $\|\cdot\|_F$ denoting the Frobenius norm. Prove mathematically that this normalization ensures scale-invariant visualization across different architectures, and explain why it is superior to standard L_2 normalization for convolutional neural networks with heterogeneous layer structures.

To demonstrate scale invariance, first examine what happens under parameter scaling transformations: For a network with scaled parameters $\theta' = c \cdot \theta$, the filter-normalized direction becomes:

$$\tilde{d}' = \frac{d}{\|d\|_{\text{filter}}} \cdot \|c \cdot \theta\|_{\text{filter}} = c \cdot \tilde{d}$$

When visualizing the loss landscape at $\theta' + \epsilon \cdot \tilde{d}'$, this is equivalent to:

$$L(\theta' + \epsilon \cdot \tilde{d}') = L(c \cdot \theta + \epsilon \cdot c \cdot \tilde{d}) = L(c(\theta + \epsilon \cdot \tilde{d}))$$

Since neural network loss functions are scale-invariant ($L(c \cdot \theta) = L(\theta)$), the resulting visualization contours remain identical despite the parameter scaling. Filter normalization is superior to L_2 normalization for CNNs because:

(a) **Layer Balance**: L_2 normalization biases perturbations toward layers with more parameters, creating distorted visualizations where:

$$\frac{\|\tilde{d}_j^l\|_F^{L2}}{\|\theta_j^l\|_F} \propto \frac{\|d_j^l\|_F}{\|\theta_j^l\|_F} \cdot \frac{\|\theta\|_2}{\|d\|_2}$$

This ratio varies across layers, causing inconsistent perturbation effects.

(b) **Architectural Invariance**: Filter normalization ensures consistent relative perturbation magnitude across filters:

$$\frac{\|\tilde{d}_j^l\|_F}{\|\theta_j^l\|_F} = \frac{\|d_j^l\|_F}{\|d\|_{\text{filter}}} \cdot \frac{\|\theta\|_{\text{filter}}}{\|\theta_j^l\|_F}$$

This preserves the network's architectural inductive bias during visualization.

(c) **Normalized Comparison**: For networks with identical architectures but different scales ($\theta_1 = \theta$, $\theta_2 = \alpha \cdot \theta$), filter normalization produces identical loss contours, while L_2 normalization would create artificially smoother landscapes for the larger network, obscuring the true geometric properties.